Description of IMG metatranscriptome data file.

RNA-seq paired-end reads are assembled by megahit (with options --k-list 23,43,63,83,103,123), and the resulting assemblies are annotated by the IMG/ER annotation pipeline. RNA-seq reads are mapped to the assemblies by bbmap in the bbtools package (with the ambiguous=random option). The RNA-seq mapping bam file is splitted into two bam files by samtools: one includes only the forward stranded reads (with samtools view flags of “-f 128 -F 16” and “ -f 80”); the other includes only the reverse stranded reads (with samtools view flags of “ -f 144” and “ -f 64 -F 16”). The read counting results for both strands are generated by the coverage function in the bedtools package with their corresponding bam files and a gff file including the gene predictions by IMG annotation.

IMG provides expression values and read counts for gene features predicted on the contigs, be it self-assembly of metatranscriptome or another dataset to which the metatranscriptome reads were mapped. Expression values are computed as mean and median per-base coverage of the sequence within the coordinates of the feature.

Since JGI generally generates stranded libraries, expression values and read counts for two strands are computed and reported separately. These values are NOT expression values and read counts of direct and reverse strand of the contig; instead these are expression values and read counts of the predicted feature (i. e. reads generated for the same strand on which the feature was predicted) and of the opposite strand of the predicted feature. Essentially this "expected" read coverage (in a sense of being generated from the strand that we expect to be expressed) and "unexpected" read coverage (i. e. generated from the strand that we did not expect to be expressed based on structural annotation of the sequence). For obvious reasons, some of the "unexpected" coverage is the result of imperfect structural annotation, which is not uncommon for short contigs in metaT self assembly.

Specific columns in the file:

img_gene_oid - gene_oid of the gene for which expression is counted

img_scaffold_oid - scaffold/contig id on which the gene has been predicted

locus_tag - another gene id of the gene for which expression is counted; this is included because all genomes and some metagenomes and metatranscriptomes used as references have both gene oids and locus tags, while others don’t

scaffold_accession - another identifier of scaffold/contig on which the gene has been predicted

strand - strand on which the gene has been predicted

locus_type - type of the gene; for example CDS (protein-coding gene), tRNA, rRNA, tmRNA, etc.

length - length of the gene for which expression is counted

reads_cnt - number of reads mapped on the same strand as predicted gene within the coordinates of the gene

mean - mean expression of the predicted gene, i. e. mean per-base coverage of the strand on which the gene was predicted within the coordinates of the predicted gene

median - median expression of the predicted gene, i. e. median per-base coverage of the strand on which the gene was predicted within the coordinates of the predicted gene

stdev - standard deviation of the expression of the predicted gene
reads_cntA - number of reads mapped to the opposite strand of the predicted gene within the coordinates of the gene
meanA - mean expression of the opposite strand of the predicted gene, i. e. mean per-base coverage of the strand opposite to that on which the gene was predicted within the coordinates of the predicted gene
medianA - median expression of the opposite strand of the predicted gene, i. e. median per-base coverage of the strand opposite to that on which the gene was predicted within the coordinates of the predicted gene
stdevA - standard deviation of the expression of the opposite strand of the predicted gene